CascadeDebate: Multi-Agent Deliberation for Cost-Aware LLM Cascades

Raeyoung Chang[*]
Sogang University
icanry@sogang.ac.kr

Dongwook Kwon[*]
Suresoft Technologies
dwkwon@suresofttech.com

Jisoo Lee[*]
Seoul National University
sally66890@snu.ac.kr

Nikhil Verma[†]
LG Electronics, Toronto AI Lab
nikhil.verma@lge.com

Abstract

Cascaded LLM systems coordinate models of varying sizes with human experts to balance accuracy, cost, and abstention under uncertainty. However, single-model tiers at each stage falter on ambiguous queries, triggering premature escalations to costlier models or experts due to under-confidence and inefficient compute scaling. CascadeDebate addresses this critical gap by inserting multi-agent deliberation directly at each tier's escalation boundary. Confidence-based routers activate lightweight agent ensembles only for uncertain cases, enabling consensus-driven resolution of ambiguities internally, without invoking higher-cost upgrades. Our unified architecture alternates single-model inference with selective multi-agent deliberation across model scales, culminating in human experts as final fallback. This design scales test-time compute dynamically to query difficulty. Across five benchmarks spanning science, medicine, and general knowledge, CascadeDebate outperforms strong single-model cascades and standalone multi-agent systems by up to 26.75%. An online threshold optimizer proves essential, boosting accuracy 20.98–52.33% relative improvement over fixed policies and enabling elastic adaptation to real-world distributions.

1 Introduction

Large language models (LLMs) have demonstrated remarkable proficiency across diverse benchmarks, spanning scientific question answering to medical diagnosis tasks (Hendrycks et al., 2021; Pal et al., 2022; Singhal et al., 2023). Despite these advances, real-world deployment demands far more than raw predictive capability. The systems must balance high accuracy against substantial inference costs, particularly as applications scale to production environments (Chen et al., 2024). High-capacity models deliver superior performance, but incur prohibitive computational overhead during training and inference (Yue et al., 2024). Compact alternatives promise efficiency, yet sacrifice reliability, remaining prone to hallucinations and breakdowns in multi-step reasoning (Wei et al., 2022; Ji et al., 2023).

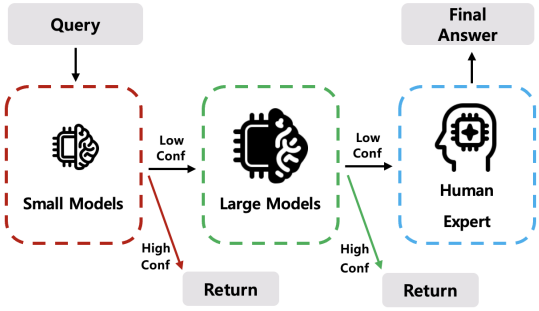


Figure 1: **Adaptive Cascade Framework Overview.** Current systems route queries through a hierarchy of solvers: from small models to large models, finally to human experts. High-confidence answers commit immediately at each stage; uncertain cases escalate to balance cost and accuracy.

As shown in Figure 1, current approaches mitigate this tension through cascaded pipelines and routing mechanisms (Aggarwal et al., 2024; Ong et al., 2025). These systems route queries from smaller models to larger ones based on confidence or acceptance criteria, deferring only when uncertainty signals potential failure. While effective for gross cost-accuracy trade-offs, such cascades rely on single-model decisions at each tier. Ambiguous inputs thus trigger premature escalations, propagating overconfident errors without intra-tier correction, and wasting compute on unnecessary upgrades (Fanconi and van der Schaar, 2025). Multi-agent frameworks offer a compelling counterpoint. They harness collective intelligence through debate, consensus, or role specialization to boost reasoning depth (Du et al., 2024; junyou li et al.,

[*]Equal Contribution.
[†]Project Lead and corresponding author.

2024; Li et al., 2023). Agents deliberate collaboratively, surfacing diverse perspectives that resolve ambiguities a lone model might mishandle. Yet these setups typically operate in isolation, as standalone enhancers rather than embedded components within cost-aware hierarchies (Wang et al., 2025). Standalone multi-agent systems overlook selective activation: deliberation excels on hard cases, but proves redundant and costly for straightforward queries.

This gap calls for a hybrid approach that embeds multi-agent deliberation within cascades, triggered only at escalation boundaries. Instead of routing uncertain queries from a single model directly to expensive larger tiers, we position agent consensus as an intermediate step. Such intra-tier scaling harnesses emergent cooperation. It self-corrects for ambiguities before any escalation occurs. This preserves efficiency while strengthening base models. Agents supply the diverse viewpoints missing from lone models, exactly where current cascades squander compute through premature deferrals. We present CascadeDebate, a unified architecture alternating single-model inference with multi-agent deliberation across scales (shown in Figure 2). Human experts serve as final fallback. Confidence-based routers control progression. High-confidence single-model outputs commit right away. Marginal cases activate lightweight agent ensembles for consensus refinement. Only unresolved uncertainty advances to larger tiers or humans.

This design embodies test-time compute scaling tailored to query difficulty where deliberation depth emerges dynamically from need rather than fixed policy. To ensure adaptability without manual tuning, we introduce an online optimizer that refines escalation thresholds from streaming human feedback. The mechanism learns task-specific boundaries online, minimizing token costs while maximizing accuracy, transforming the cascade into an elastic reasoner responsive to real-world query distributions. Our contributions advance this vision:

- A unified cascade architecture interposing multi-agent deliberation at each tier's escalation boundary, with human-in-the-loop as ultimate recourse.
- An online threshold learner balancing accuracy against computation, enabling continual refinement from human signals.
- Empirical validation showing that adaptive stage selection outperforms single-model cascades and standalone agents, resolving ambiguities internally to reduce expert load substantially.

This architecture not only bridges cascades and multi-agent paradigms but deploys practical human-AI collaboration at scale. Subsequent sections detail the architecture (Sec. 3), experiment setup (Sec. 4.1), and rigorous evaluation (Sec. 4.2).

2 Related Work

2.1 Cost-Aware Routing and LLM Cascades

Adaptive inference systems optimize cost-accuracy trade-offs by routing queries across model tiers (Chen et al., 2024; Aggarwal et al., 2024). Cascaded architectures prioritize lightweight models and escalate only when acceptance criteria are not met (Ong et al., 2025; Ding et al., 2024; Yue et al., 2024). These frameworks typically use single-model inference at each decision point, which makes routing sensitive to noisy confidence estimates and calibration error (Xiong et al., 2024; Fanconi and van der Schaar, 2025). CascadeDebate addresses this limitation by adding selective intra-tier computation that targets marginal cases before inter-tier escalation.

2.2 Multi-Agent Deliberation and Consensus

Multi-agent systems improve reasoning robustness through debate, critique, and role specialization, followed by aggregation (Du et al., 2024; Lee et al., 2025; Li et al., 2023; junyou li et al., 2024). These mechanisms reduce single-model variance and expose failure modes that isolated inference may miss (Zhang et al., 2024; Wang et al., 2025). Prior work also reduces coordination overhead through sparse interaction structures or structured roles (Sun et al., 2025; Chen et al., 2025). However, deliberation is typically deployed as a standalone enhancement rather than as a selective operator integrated with cascade routing decisions. CascadeDebate integrates deliberation at escalation boundaries to align additional computation with uncertainty.

2.3 Online Threshold Learning

LLM confidence estimates are poorly calibrated (Xiong et al., 2024), making static routing thresholds brittle under distribution shifts and heterogeneous domain difficulty, motivating adaptive deferral policies (Fanconi and van der Schaar, 2025).

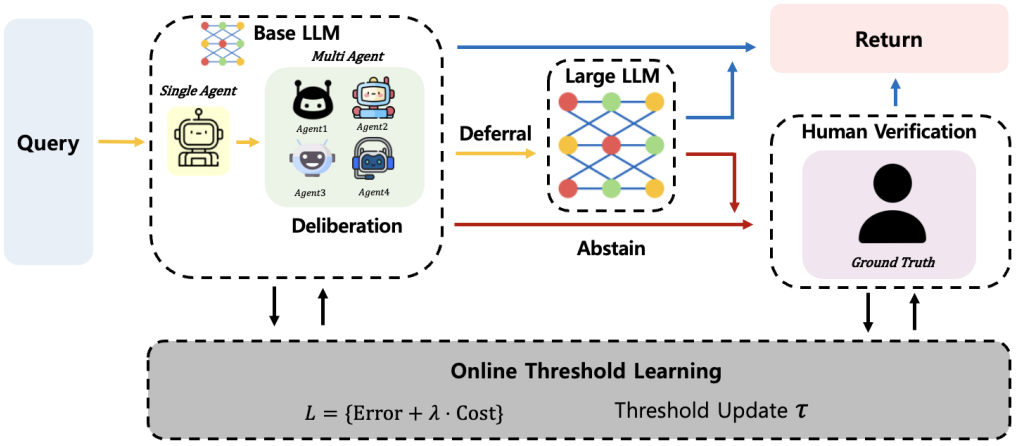


Figure 2: **CascadeDebate Architecture.** The unified framework alternates single-model inference with multi-agent deliberation across base and large scales, culminating in human experts as final fallback. Confidence-based routers activate lightweight agent ensembles only for marginal cases at each escalation boundary, resolving ambiguities via intra-tier consensus before deferring to costlier tiers. The online threshold optimizer continuously refines deferral boundaries $\boldsymbol{\tau}$ from human feedback, enabling elastic adaptation to real-world query distributions while balancing accuracy against token and expert costs.

Unified formulations connect routing and cascading to support learning deferral behavior instead of manual threshold selection (Dekoninck et al., 2024; Shen et al., 2025). CascadeDebate follows this direction by updating escalation thresholds online from streaming human feedback. The resulting policy adapts stage utilization to the observed workload while preserving explicit control over the cost-accuracy trade-off.

3 Methodology

CascadeDebate extends cascaded decision systems by embedding multi-agent deliberation at escalation boundaries between model tiers. We denote the cascade as $C = \{S_1, S_2, \ldots, S_K\}$, where each stage S_k processes input $x \in X$ to produce answer $\hat{y}_k \in Y$, confidence $\Phi_k(x) \in [0, 1]$, and uncertainty $\Xi_k(x) \in [0, \infty)$, following terminology of (Fanconi and van der Schaar, 2025). At each stage, progression follows the deferral policy: accept $\hat{y}_k$ if $\Phi_k(x) > \tau_k^d$ (confidence exceeds deferral threshold); abstain to humans if $\Xi_k(x) > \tau_k^a$; otherwise defer to S_{k+1}. For basic cascade we follow two-tier model architecture $M_{\text{base}} \to M_{\text{large}}$, but it can be generalized to arbitrary depth K.

3.1 CascadeDebate architecture

It alternates single-model inference S_{single} with multi-agent deliberation S_{multi} across model scales. Formally, odd stages use single inference:

$$S_{\text{single}}(x; \mathcal{M}_i) = \big(\hat{y}, \Phi_i(x), \Xi_i(x)\big),$$

where $\mathcal{M}_i \in \{\mathcal{M}_{\text{base}}, \mathcal{M}_{\text{large}}\}$ denotes model scale i. Even stages activate deliberation only on marginal confidence:

$$S_{\text{multi}}(x; \mathcal{M}_i, N) = \text{Consensus}\big(\{\mathcal{M}_i(x; r_j)\}_{j=1}^{N}\big),$$

where r_j are distinct role prompts and Consensus aggregates via majority vote with confidence $\Phi_{\text{multi}}(x) = \mathbb{P}(\text{agreement})$.

This design scales test-time compute intra-tier before inter-tier escalation. Marginal cases trigger S_{multi} to resolve ambiguities via emergent cooperation, delaying costly upgrades. The architecture naturally extends to additional model scales by repeating the single→multi pattern.

3.2 Confidence Estimation

CascadeDebate employs complementary confidence signals tailored to stage type, followed by Bayesian calibration. For single-model stages S_{single}, we extract $\Phi_k(x)$ via surrogate token probability (Kadavath et al., 2022), capturing model self-assessment of answer quality. For multi-agent stages S_{multi}, confidence reflects ensemble agreement: $\Phi_k(x) = \phi_k^{\text{agree}}$ of N agents $\mathcal{M}_i(x; r_j)$. The majority vote prediction and the agreement confidence are calculated as:

$$\hat{y}_k^{\text{mv}} = \arg\max_a \sum_{j=1}^{N} \mathbb{1}[\hat{y}_k^{(j)} = a] \tag{1}$$

$$\phi_k^{\text{agree}} = \frac{|\{j : \hat{y}_k^{(j)} = \hat{y}_k^{\text{mv}}\}|}{N} \tag{2}$$

Multi-agent deliberation provides robust intra-tier signal where high agreement rate resolved ambiguity without escalation.

Both signals undergo Bayesian logistic regression calibration, fitted on 100 held-out samples per model scale $\mathcal{M}_i$. Calibration ensures $\Phi_k(x) \approx \mathbb{P}(\hat{y}_k = y^*|x)$ across operating ranges.

3.3 Online Threshold Learning

Thresholds $\boldsymbol{\tau} = \{\tau_k^d, \tau_k^a\}_k$ govern routing, i.e. to accept if $\Phi_k(x) > \tau_k^d$, abstain if $\Xi_k(x) > \tau_k^a$. Following the setup of (Fanconi and van der Schaar, 2025), we parameterize it via sigmoid: $\tau_k = \sigma(\theta_k)$ with $\theta_k \in \mathbb{R}$. Soft gating enables gradient optimization: accept probability $\pi_k = \sigma(\gamma.(\Phi_k(x) - \tau_k))$ where $\gamma = 5$ controls sharpness. Stage k stopping probability chains prior deferrals: $p_k = \pi_k \prod_{i<k}(1-\pi_i)$. The multi-objective loss balances error and cost:

$$\mathcal{L}(\boldsymbol{\theta}) = (1 - \mathbb{E}[p_k \cdot \text{corr}_k]) + \lambda \cdot \mathbb{E}[p_k \cdot c_k] \quad (3)$$

where $\mathbb{E}[\cdot]$ aggregates over reached stages, $\text{corr}_k = \mathbb{1}[\hat{y}_k = y^*]$, and c_{expert} assumes perfection. Parameters update via Adam on mini-batches from replay buffer $\mathcal{B}$, accumulating human feedback post-deployment. More details in Appendix A.

4 Experiments and Results

The full cascade is

$$\begin{aligned} C = {} & S_{\text{single}}(\mathcal{M}_{\text{base}}) \rightarrow S_{\text{multi}}(\mathcal{M}_{\text{base}}) \\ & \rightarrow S_{\text{single}}(\mathcal{M}_{\text{large}}) \rightarrow S_{\text{multi}}(\mathcal{M}_{\text{large}}) \\ & \rightarrow S_{\text{human}} \end{aligned}$$

terminating early per thresholds $\boldsymbol{\tau} = \{\tau_k^d, \tau_k^a\}_k$.

4.1 Setup details

We evaluate CascadeDebate on five multiple-choice benchmarks, sampling 1,000 instances from each. These span science (ARC-Easy and ARC-Challenge (Clark et al., 2018)), general knowledge (MMLU (Hendrycks et al., 2021)), and medicine (MedQA (Jin et al., 2021); MedMCQA (Pal et al., 2022)). We use test splits for ARC and MMLU, validation splits for medical datasets, with statistics in Appendix B. We employ instruction-tuned Llama-3.2 (Dubey et al., 2024) and Qwen2.5 (Yang et al., 2024). Base models are Llama-3.2-1B-Instruct and Qwen2.5-1.5B-Instruct, while large models are Llama-3.2-3B-Instruct and Qwen2.5-3B-Instruct. All generation used $temp = 0$ during inference with maximum length of 512 tokens.

Each dataset uses four specialized role prompts per multi-agent stage (4 agents). For e.g., science tasks include *Experimental Scientist* and *Misconception Detector*; medical tasks include *Clinical Reasoning* and *Evidence-Based Medicine* (full prompts in Appendix C). All experiments run on a single NVIDIA A100 (80GB) GPU using PyTorch 2.4 and Transformers v4.44.2. Thresholds optimize via Adam (batch size 10, $\eta_\tau = 0.05$), initializing to $\tau_k^d = 0.6$. Hyperparameters are $\lambda = 10^{-7}$ (cost-accuracy trade-off), $\rho = 5.0$ (output-to-input price ratio), and $c_{\text{expert}} = 10.0$.

4.2 Results & Observations

CascadeDebate answers two key questions across five benchmarks:

1. Does embedding multi-agent deliberation at escalation boundaries outperform individual-stage baselines?

2. How does the system allocate compute across capability tiers?

We compare CascadeDebate against single-model, multi-agent, and standard cascade baselines across model scales (Table 1). A consistent trend is that multi-agent variants outperform their single-model counterparts, and larger models further improve accuracy within each dataset. Standard cascades that route from $S_{\text{single}}(\mathcal{M}_{\text{base}}) \rightarrow S_{\text{single}}(\mathcal{M}_{\text{large}}) \rightarrow S_{\text{human}}$ already surpass any individual stage, confirming the value of confidence-based deferral even without intra-tier deliberation.

Superior accuracy at controlled cost. CascadeDebate achieves the best accuracy on all tasks with Llama-3.2, improving over the strongest baseline by 1.43-18.24 percentage points (pp) (e.g., MedQA: 86.44% vs. 64.00% for $S_{\text{multi}}(\mathcal{M}_{\text{large}})$ and 68.20% for the standard cascade). For Qwen2.5, CascadeDebate matches or exceeds the standard cascade on three of five datasets and yields especially large gains on the medical benchmarks (e.g., MedQA: 75.22% vs. 52.70%). These patterns show that inserting deliberation at escalation boundaries systematically boosts accuracy beyond what can be obtained from scaling model size or naïve cascades alone.

Effect of model scale and deliberation. Within each backbone, moving from $S_{\text{single}}(\mathcal{M}_{\text{base}})$ to $S_{\text{single}}(\mathcal{M}_{\text{large}})$ yields substantial gains (e.g., Llama-3.2 MMLU: 44.22% $\rightarrow$ 62.44%), and adding multi-agent deliberation on top of a fixed

Table 1: **CascadeDebate vs. Baselines.** Accuracy (%) of Llama-3.2-Instruct (1B/3B) and Qwen2.5-Instruct (1.5B/3B) across single-model ($S_{\text{single}}(\mathcal{M}_{\text{base/large}})$), multi-agent ($S_{\text{multi}}(\mathcal{M}_{\text{base/large}})$), standard cascade ($S_{\text{single}}(\mathcal{M}_{\text{base}}) \rightarrow S_{\text{single}}(\mathcal{M}_{\text{large}}) \rightarrow S_{\text{human}}$), and CascadeDebate ($S_{\text{single}}(\mathcal{M}_{\text{base}}) \rightarrow S_{\text{multi}}(\mathcal{M}_{\text{base}}) \rightarrow S_{\text{single}}(\mathcal{M}_{\text{large}}) \rightarrow S_{\text{multi}}(\mathcal{M}_{\text{large}}) \rightarrow S_{\text{human}}$) systems. Best in **bold**.

Dataset	Baselines				Cascade	
	$S_{\text{single}}(\mathcal{M}_{\text{base}})$	$S_{\text{multi}}(\mathcal{M}_{\text{base}})$	$S_{\text{single}}(\mathcal{M}_{\text{large}})$	$S_{\text{multi}}(\mathcal{M}_{\text{large}})$	**Standard**	**Ours**
Llama-3.2 (Base: 1B, Large: 3B)						
ARC-Easy	68.56	73.89	89.67	91.00	93.90	$\mathbf{95.33}_{\pm 0.70}$
ARC-Challenge	50.67	54.78	78.33	81.67	84.30	$\mathbf{92.89}_{\pm 0.86}$
MMLU	44.22	47.56	62.44	64.78	67.70	$\mathbf{82.67}_{\pm 1.26}$
MedQA	34.22	35.11	60.89	64.00	68.20	$\mathbf{86.44}_{\pm 1.14}$
MedMCQA	36.11	41.78	52.67	55.33	64.70	$\mathbf{76.33}_{\pm 1.42}$
Qwen2.5 (Base: 1.5B, Large: 3B)						
ARC-Easy	83.44	88.44	93.44	94.33	**96.20**	$92.00_{\pm 0.01}$
ARC-Challenge	71.89	74.67	82.78	84.44	**89.80**	$85.78_{\pm 0.01}$
MMLU	58.78	58.44	65.00	67.11	73.20	$\mathbf{78.00}_{\pm 0.01}$
MedQA	34.44	36.89	40.89	41.44	52.70	$\mathbf{75.22}_{\pm 0.01}$
MedMCQA	37.33	42.00	46.44	46.67	58.70	$\mathbf{69.89}_{\pm 0.02}$

scale yields further improvements (62.44% $\rightarrow$ 64.78% for $S_{\text{multi}}(\mathcal{M}_{\text{large}})$). However, CascadeDebate consistently outperforms both single and multi-agent large models (e.g., Llama-3.2 MedMCQA: 55.33% vs. 76.33%), indicating that selective intra-tier cooperation is more effective than uniformly applying multi-agent methods at a single scale.

Benefit over standard cascades. Standard cascades already close much of the gap between base and large models (e.g., Qwen2.5 ARC-Challenge: 71.89% $\rightarrow$ 89.80%), demonstrating that confidence-based routing is a strong baseline. CascadeDebate nonetheless delivers additional gains, most notably on harder and medical benchmarks where ambiguity is common. These improvements support our hypothesis that many escalations can be resolved by intra-tier debate rather than blindly deferring to larger or human solvers.

Implications for compute allocation. Detailed cost-accuracy curves are shown in Fig. 3. CascadeDebate systematically converts additional compute into disproportionate accuracy gains across domains. On relatively easier ARC-Easy (Llama-3.2), CascadeDebate delivers +26.78pp (68.56% $\rightarrow$ 95.33%) at only 12.79$\times$ Single Base cost, surpassing even $S_{\text{multi}}(\mathcal{M}_{\text{large}})$ (91.00%) while selectively activating lightweight base deliberation. For complex reasoning in ARC-Challenge, cost rises to 15.62$\times$ but yields massive +42.22pp (50.67% $\rightarrow$ 92.89%), 1.62$\times$ the gain of standard cascade (76.78%). The most compelling results appear on medical benchmarks: MedQA achieves +52.22pp (34.22% $\rightarrow$ 86.44%) at 16.66$\times$ cost; MedMCQA gains +40.22pp (36.11% $\rightarrow$ 76.33%) at 13.71$\times$.

These trends confirm the core hypothesis: intra-tier multi-agent deliberation resolves good amount of escalation cases internally, concentrating expensive large-model and S_{human} compute precisely where single models fail catastrophically. The system processes straightforward queries efficiently via $S_{\text{single}}(\mathcal{M}_{\text{base}})$ while reserving intensive reasoning for true edge cases, achieving cost-accuracy trade-offs. Similar patterns hold for Qwen2.5 (see Appendix D, Fig. 4).

5 Discussion

CascadeDebate demonstrates that carefully orchestrated small models can surpass significantly larger standalone systems. Here, we examine the mechanisms driving this performance, the role of adaptive thresholding, practical implications, and key limitations.

Inference-time compute trumps parametric scaling. Our 1B-parameter online cascade consistently outperforms 3B standalone models by 14.56-23.66 pp across challenging benchmarks (ARC-Challenge: 92.89% vs. 78.33%; MedMCQA: 76.33% vs. 52.67%). This challenges conventional scaling laws by showing that inference-time compute, through selective multi-agent deliberation, can generate more effective reasoning than additional parameters alone. Selective multi-agent deliberation rectifies marginal uncertainties internally, enabling compact models to overcome paramet-

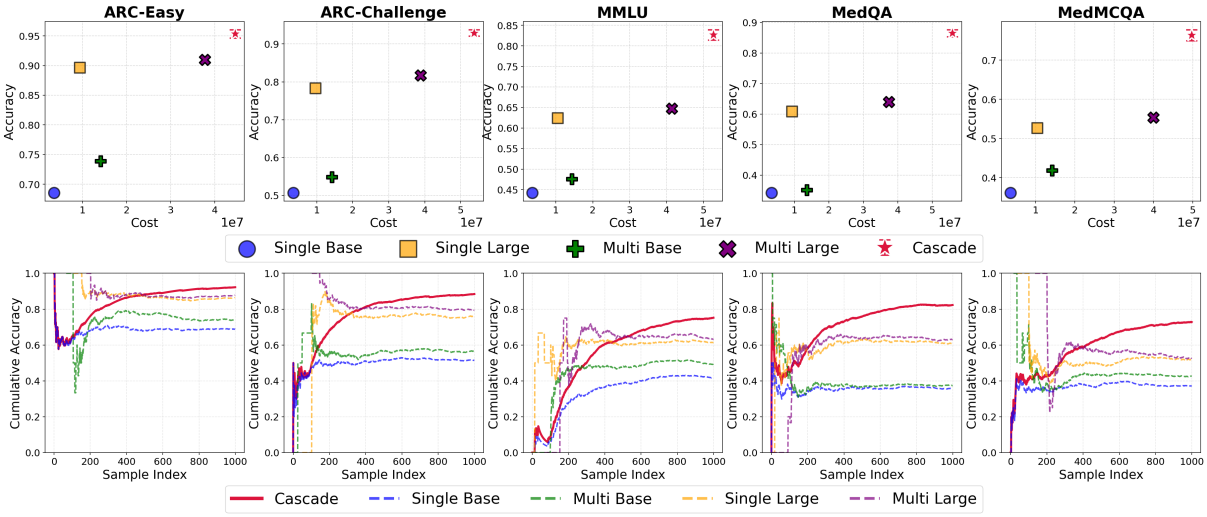


Figure 3: **CascadeDebate Cost-Accuracy Pareto Frontier. Top:** Cost-accuracy curves across five benchmarks using Llama-3.2-Instruct model. CascadeDebate (red $\star$) dominates the Pareto frontier, matching or exceeding $S_{\text{multi}}(\mathcal{M}_{\text{large}})$ accuracy at 20-35% lower token cost. **Bottom:** Online threshold adaptation. CascadeDebate (solid red) rapidly surpasses fixed-threshold cascade (dashed red) and static baselines as the optimizer calibrates $\boldsymbol{\tau}_k$ to task difficulty over 1,000 samples.

ric limitations through targeted test-time compute rather than uniform redundancy.

Online adaptation is critical. Fixed thresholds ($\tau_k = 0.6$) perform reasonably but plateau between base and large model performance, as a single cutoff proves suboptimal across domains. Our online optimizer transforms this limitation, learning task-specific boundaries that balance frugality and safety by processing 44.40% of ARC-Easy queries at the first stage while aggressively escalating only true edge cases in medical domains. Refer to Appendix E (Table 5) for comparison of fixed to learned threshold.

Elastic resource allocation. The architecture exhibits confidence-driven elasticity. It incurs 12.79$\times$ Single Base cost on ARC-Easy versus 15.62$\times$ on ARC-Challenge, yielding accuracy gains of 26.78pp to 42.22pp respectively. This adaptability delivers Pareto-superior cost-accuracy tradeoffs, concentrating expensive human and large-model compute precisely where single models fail catastrophically. We report stage-wise distributions in Appendix E (Fig. 6).

Quantified latency analysis Despite sequential staging, CascadeDebate achieves strong accuracy gains with minimal latency overhead vs multi-agent baselines. Sequential latency remains a noted limitation (Sections 5, Limitations). On ARC-Challenge (Qwen2.5, A100), it adds only +0.1s (4.4%) over $S_{multi}(M_{large})$ while improving accuracy as shown in Table 2. CascadeDebate's 2.45s latency remains minimal relative to multi-agent vs their single-agent baselines, validating its efficiency for latency-tolerant applications.

Table 2: Model performance on ARC-Challenge with mean latency.

Model	Mean Latency (sec)	Accuracy (%)
$S_{\text{single}}(M_{\text{base}})$	0.38	71.89
$S_{\text{multi}}(M_{\text{base}})$	1.45	74.67
$S_{\text{single}}(M_{\text{large}})$	0.63	82.78
$S_{\text{multi}}(M_{\text{large}})$	2.34	84.44
CascadeDebate	**2.45**	**85.78**

Human expert cost modeling assumes perfect accuracy, which may overestimate final-stage performance in practice. Future work will explore automated role discovery, test-time distillation of consensus reasoning, and deployment in long-context production workloads. We will also extend evaluation to larger-scale models (e.g., 70B+) and heterogeneous ensembles to better understand scaling behavior of selective intra-tier deliberation. Finally, we will broaden evaluation beyond multiple-choice QA to open-ended generation settings and investigate more diverse agent configurations and adaptive role assignments within our Bayesian-calibrated routing framework.

6 Conclusion

CascadeDebate embeds multi-agent deliberation at cascade escalation boundaries, alternating single-

model inference with selective agent ensembles across scales, with human experts as final fallback. Confidence-based routers activate consensus only for marginal cases, resolving ambiguities internally before costly escalation. Our online threshold optimizer transforms static cascades into elastic reasoners, adapting to query distributions while balancing accuracy against compute costs. This inference-time scaling outperforms traditional parametric approaches across diverse domains. By concentrating expensive compute where single models fail most catastrophically, CascadeDebate delivers superior cost-accuracy tradeoffs with human-like cognitive elasticity-lightweight for routine queries and intensive deliberation for edge cases. Future work will explore parallel execution, automated role discovery, and extension to open-ended generation.

Limitations

We acknowledge three limitations. First, latency arises from the sequential nature that introduces cumulative delays that exceed single-pass models. Second, error propagation occurs when miscalibrated base models prematurely accept incorrect answers and prevent the necessary escalation. Third, agent homogeneity limits diversity compared to heterogeneous ensembles as we rely solely on role-prompting within a single model family.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00143911, AI Excellence Global Innovative Leader Education Program). This work was also supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00995, Automated Reliable Source Code Generation from Natural Language Descriptions).

References

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, and 1 others. 2024. Automix: Automatically mixing language models. *Advances in Neural Information Processing Systems*, 37:131000–131034.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*. Featured Certification.

Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. 2025. Multi-agent evolve: Llm self-improve through co-evolution. *arXiv preprint arXiv:2510.23595*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. 2024. A unified approach to routing and cascading for llms. *arXiv preprint arXiv:2410.10347*.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid LLM: Cost-efficient and quality-aware query routing. In *The Twelfth International Conference on Learning Representations*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Claudio Fanconi and Mihaela van der Schaar. 2025. Cascaded language models for cost-effective human-ai decision-making. *arXiv preprint arXiv:2506.11887*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

junyou li, Qin Zhang, Yangbin Yu, QIANG FU, and Deheng Ye. 2024. More agents is all you need. *Transactions on Machine Learning Research*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jisoo Lee, Raeyoung Chang, Dongwook Kwon, Harmanpreet Singh, and Nikhil Verma. 2025. Gemmas: Graph-based evaluation metrics for multi agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1522–1532.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Yuanzhe Shen, Yide Liu, Zisu Huang, Ruicheng Yin, Xiaoqing Zheng, and Xuan-Jing Huang. 2025. Sater: A self-aware and token-efficient approach to routing and cascading. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10526–10540.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. 2025. Cortexdebate: Debating sparsely and equally for multi-agent debate. In *ACL (Findings)*, pages 9503–9523.

Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2024. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *The Twelfth International Conference on Learning Representations*.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. 2024. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

A Online Learning Algorithm

The confidence $\Phi_k(x)$ and uncertainty $\Xi_k(x)$ metrics used in *CascadeDebate* are designed to provide complementary signals for intra-tier and inter-tier escalation. For single-model stages, we utilize the *Surrogate Token Probability* (STP), whereas for multi-agent stages, Φ_k is defined as the *Agreement Rate* among N agents. Both are naturally bounded as $\Phi_k(x) \in [0, 1]$, representing a probabilistic estimate of correctness. In contrast, the uncertainty metric $\Xi_k(x) \in [0, \infty)$ quantifies the "unknown unknowns" or epistemic uncertainty, typically computed using the logit entropy of the base model or the variance across agent outputs. While clear queries yield values near 0, highly ambiguous inputs result in high entropy that theoretically approaches infinity, triggering an escalation to human experts. To ensure these raw signals are reliable, we apply *Bayesian Logistic Regression* fitted on a held-out calibration set. This process maps raw scores to empirical accuracy $P(\hat{y}_k = y^*|x)$, correcting for the typical over-confidence observed in large language models.

Since the decision to accept or defer at each stage is discrete (based on the indicator function $\mathbb{I}[\phi_k \geq \tau_k]$), the standard gradient descent cannot be applied directly. To enable end-to-end differentiability, we employ a soft relaxation technique during the backward pass.

We parameterize each threshold τ_k as a scalar $\theta_k \in \mathbb{R}$ such that $\tau_k = \sigma(\theta_k)$, ensuring that the threshold remains within $(0, 1)$. During optimization, we approximate the "pass probability" (probability of deferring to the next stage) using a shifted sigmoid function:

$$\pi_k \approx \sigma\left(\kappa \cdot (\tau_k - \phi_k)\right) \tag{4}$$

where κ is a temperature scaling factor (set to 5.0) that controls the sharpness of the decision boundary. Using this soft probability, the expected cumulative cost and expected error become differentiable functions of θ_k.

The Algorithm 1 details the inference and update loop. We maintain a replay buffer $\mathcal{B}$ to store recent query tuples $(\phi_{1:4}, y_{\text{true}}, c_{1:4})$. In update steps, we sample a mini-batch from $\mathcal{B}$ and perform a gradient update on $\{\theta_k\}$ to minimize compound loss $\mathcal{L}$ using the Adam optimizer.

Algorithm 1 Five-Stage Cascade with Online Threshold Learning

Require: Query x, thresholds $\{\tau_k\}_{k=1}^{4}$, models $\mathcal{M}_{\text{base}}, \mathcal{M}_{\text{large}}$, role prompts $\{r_j\}_{j=1}^{N}$
Ensure: Answer $\hat{y}$, cost C

```
C ← 0
for k = 1 to 4 do
    Generate ŷ_k using stage S_k configuration
    Compute confidence φ_k
    C ← C + c_k
    if φ_k ≥ τ_k then
        return ŷ_k, C {Accept at stage k}
    end if
end for
return y_expert, C + c_expert {Defer to human}
// Online update (if enabled):
Append (φ_{1:4}, 𝟙[ŷ_k=y], c_{1:4}) to B
Update {θ_k} on mini-batch from B
```

B Dataset Statistics

Table 3 provides a detailed summary of the five benchmarks used in our evaluation. To ensure a consistent computational budget across all domains, we randomly sampled N=1,000 instances from each dataset.

For ARC and MMLU, we utilized the official test splits. However, for medical datasets (MedQA and MedMCQA), the test set labels are often withheld for leaderboard competitions; therefore, we performed evaluations on the validation splits.

Benchmarks Descriptions.

- ARC-Easy & Challenge (Clark et al., 2018): Sourced from grade-school science exams. Easy questions are solvable via retrieval, whereas Challenge questions contain adversarial choices requiring multi-hop reasoning.
- MMLU (Hendrycks et al., 2021): A massive multitask benchmark covering 57 subjects (e.g., math, history, law) ranging from elementary to professional levels.
- MedQA (Jin et al., 2021): Derived from the US Medical Licensing Examination (USMLE), which requires extensive professional knowledge and clinical reasoning.
- MedMCQA (Pal et al., 2022): Collected from Indian medical entrance exams (AIIMS,

Table 3: Summary of datasets. We use 1,000 randomly sampled instances from each dataset for consistent evaluation.

Dataset	Domain	Split	Choices	Samples
ARC-Easy	Science	Test	3–5	1,000
ARC-Challenge	Science	Test	3–5	1,000
MMLU	General	Test	4	1,000
MedQA	Medical	Validation	5	1,000
MedMCQA	Medical	Validation	4	1,000

Table 4: Domain-specific agent roles used in the multi-agent stages.

Dataset	Agent Roles
ARC	Causal Chain Specialist, Misconception Detector, Experimental Scientist, Reviewer
MedQA	Clinical Reasoning Expert, Evidence-Based Medicine Specialist, Option Eliminator, Safety Reviewer
MedMCQA	High-Yield Facts Expert, Clinical Pattern Recognizer, Strategic Eliminator, Exam Strategist
MMLU	First-Principles Thinker, Context Analyst, Consistency Checker, Textbook Expert

NEET), featuring high-difficulty multiple-choice questions.

C Role Prompts

Each stage of multiple agents (S_2, S_4) utilizes four domain-specific role triggers to induce diverse perspectives of reasoning. Table 4 lists the specific roles assigned for each data set. At inference time, each agent is instantiated with a unique role system prompt to generate an independent response, after which their outputs are aggregated via majority voting.

D Results on Qwen2.5

To verify the generalizability of our framework in different model architectures, we extended our experiments to the Qwen2.5 family (Yang et al., 2024). Specifically, we designated `Qwen2.5-1.5B-Instruct` as the Base Model (Stages S_1, S_2) and `Qwen2.5-3B-Instruct` as the Large Model (Stages S_3, S_4).

Figure 4 (Top Row) presents the cost-accuracy trade-off on five benchmarks. Consistent with the main results using Llama-3.2, the Qwen-based Cascade system (marked with a red star $\star$) successfully identifies the optimal operating point.

- **Superiority over Static Baselines:** The cascade consistently outperforms the *Single Base* and *multi-agent-base* models in accuracy by a significant margin. Crucially, it achieves performance comparable to or exceeding the computationally expensive *Multi-Agent Large* baseline while consuming significantly fewer tokens.

- **Robustness across Domains:** Whether in scientific reasoning (ARC) or medical knowledge (MedQA), the framework maintains a high position on the Pareto frontier.

The learning curves (Bottom Row) confirm the stability of our optimization algorithm in the Qwen architecture. The cumulative accuracy improves rapidly in the early phase (first 100–200 samples) as the thresholds calibrate to the data set's difficulty. This indicates that our proposed loss function and STP method are robust transferable components that function reliably across different LLM families.

E Additional Experimental Analysis

In this section, we provide a deeper analysis of the cascade's internal dynamics and conduct an ablation study of the thresholding strategy. All analyzes presented here are based on the Llama-3.2 family (1B and 3B) as backbone models.

E.1 Comparison with Fixed Thresholds

To validate the need for online optimization, we analyze the performance of a static cascade where all thresholds are fixed at $\tau = 0.6$ (Figure 5).

Unlike the online approach, the *Fixed Cascade* fails to achieve the optimal Pareto frontier. As shown in the stage distribution ratios (Bottom Row), a static threshold results in a rigid allocation policy—often over-spending on simple queries or under-utilizing capable models on hard queries.

To quantify the performance gap, Table 5 compares the precision of the Fixed Threshold strategy ($\tau = 0.6$) with our Online Learning approach. The results show that the online strategy produces substantial accuracy improvements ranging from +16.1% to +26.7%.

E.2 Cascade Dynamics and Elasticity

Figure 6 shows the stage-wise distribution of query termination across benchmarks under online learned thresholds.

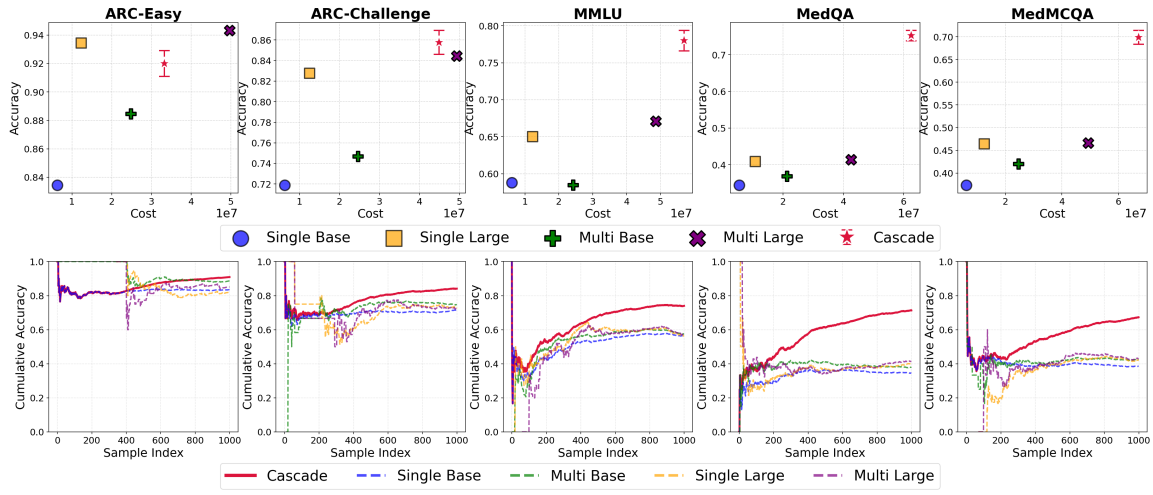


Figure 4: Generalizability Analysis using Qwen2.5 Models. The performance trends are consistent with the Llama-3.2 results. Top Row: The *Cascade* (red star $\star$) consistently occupies the optimal position on the Pareto frontier across all benchmarks. Bottom Row: The online threshold learning curve demonstrates rapid adaptation and stability, confirming that our proposed cost-aware optimization is robust regardless of the underlying backbone model.

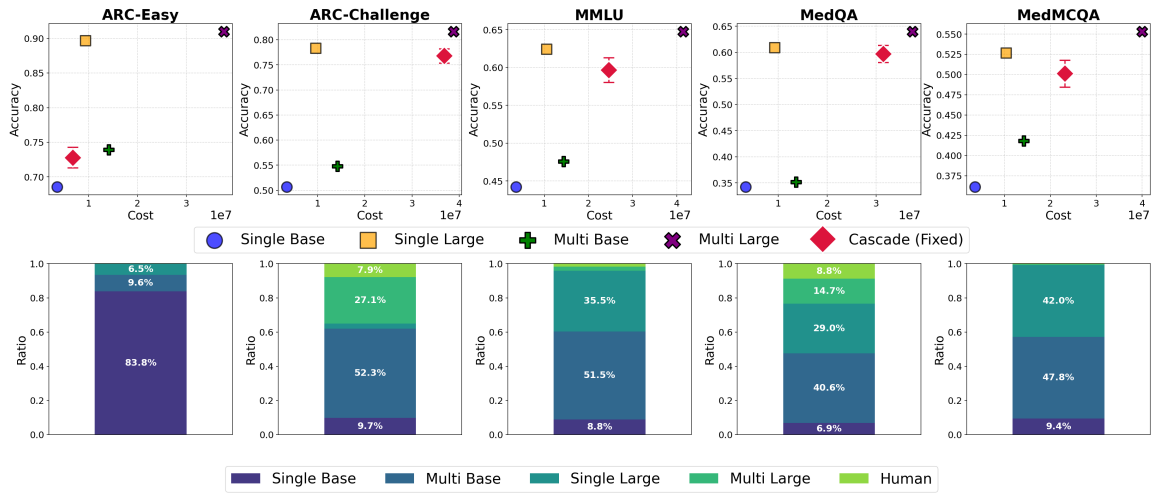


Figure 5: Performance Analysis of the Fixed Threshold Strategy on Llama-3.2 ($\tau = 0.6$). Top Row: Cost-Accuracy trade-off. The *Fixed Cascade* (red diamond $\diamondsuit$) generally improves upon the Single Base model but often falls short of the optimal Pareto frontier compared to the Online strategy. Bottom Row: Stage distribution ratios. Without dynamic adjustment, the system relies on static confidence scores, resulting in a rigid allocation of resources.

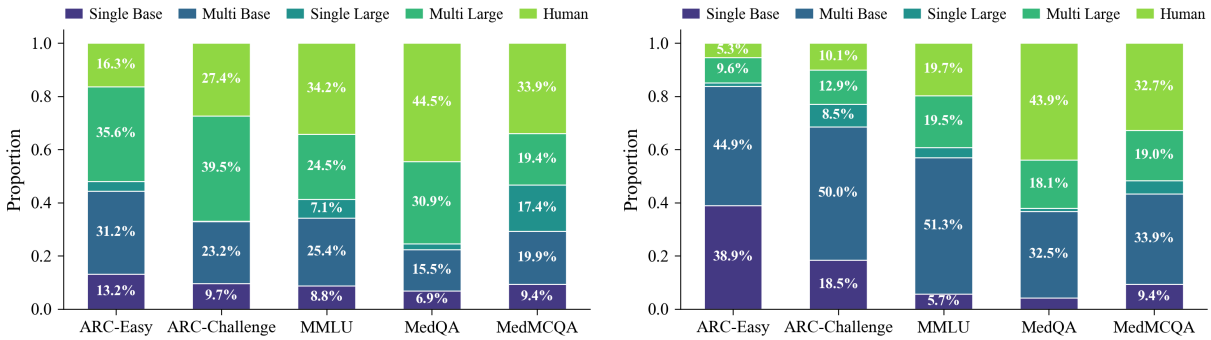
Table 5: Impact of Online Threshold Optimization. Comparison of accuracy between the Fixed Threshold strategy and the Online Learning strategy.

Dataset	Fixed Threshold	Online Learning	Improvement
ARC-Easy	72.8%	95.3%	+22.5%
ARC-Challenge	76.8%	92.9%	+16.1%
MMLU	59.7%	82.7%	+23.0%
MedQA	59.7%	86.4%	+26.7%
MedMCQA	50.1%	76.3%	+26.2%

- **Elastic Resource Allocation:** The cascade exhibits confidence-driven elasticity. On simpler benchmarks such as ARC-Easy, a larger fraction of queries terminate in earlier stages because the base model confidence more often exceeds the learned acceptance thresholds. On harder benchmarks such as MedQA, early-stage confidence more frequently falls below threshold, leading to increased deferral to later stages, including the multi-stages and the human expert stage.

F Human Involvement and Cascades

Human involvement in AI systems broadly refers to settings where people participate in an other-



(a) Llama-3.2 (b) Qwen2.5

Figure 6: Stage-wise Sample Distribution with online learned thresholds for (a) Llama-3.2 and (b) Qwen2.5. Bars show the proportion of queries that terminate at each stage under confidence-based routing. Easier benchmarks are more likely to meet the learned acceptance thresholds in earlier stages and terminate without escalation. Harder benchmarks more frequently fall below early-stage thresholds, triggering deferral to later stages including larger models and the human expert.

wise automated decision process, for example, by reviewing uncertain outputs, correcting errors, or serving as final adjudicators for unresolved or high-risk cases. A cascade is a staged architecture that routes each input through solvers with different cost-capability profiles, typically starting with cheaper models and escalating only when confidence is low or the current stage abstains (Aggarwal et al., 2024; Ong et al., 2025). The general goal is to match computation and oversight with query difficulty, so routine cases are handled efficiently, while ambiguous cases receive stronger models or human review (Fanconi and van der Schaar, 2025).